DOCUMENT RESUME

ED 393 935 TM 024 970

AUTHOR	Bennett, Randy Elliot; And Others
TITLE	A Task Type for Measuring the Representational Component of Quantitative Proficiency. GRE Board Professional Report No. 92-05P.
INSTITUTION	Educational Testing Service, Princeton, N.J.
SPONS AGENCY	Graduate Record Examinations Board, Princeton, N.J.
REPORT NO	ETS-RR-95-19
PUB DATE	Aug 95
NOTE	62p.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	*Admission (School); *Classification; College Entrance Examinations; College Faculty; *Computer Assisted Testing; Educational Assessment; Grades (Scholastic); Graduate Study; Higher Education; Mathematics Tests; Multidimensional Scaling; Pilot Projects; *Student Attitudes; *Test Construction; *Undergraduate Students
IDENTIFIERS	*Graduate Record Examinations

ABSTRACT

Two computer-based categorization tasks were developed and pilot tested. In study 1, the task asked examinees to sort mathematical word problem stems according to prototypes. Results with 9 faculty members and 107 undergraduates showed that those who sorted well tended to have higher Graduate Record Examination General Test scores and college grades than those who sorted less proficiently. Examinees generally preferred this task to multiple-choice items and felt that the task was a fairer measure of their ability to succeed in graduate school. For study 2, the task involved rating the similarity of item pairs. Five mathematics test developers and 35 undergraduate students participated, with the results analyzed by individual differences multidimensional scaling. Experts produced more scaleable ratings overall and primarily attended to two dimensions. Students used the same two dimensions, with the addition of a third. Students who rated more like experts tended to have higher admissions test scores. Examinees preferred multiple-choice questions to the rating task and felt them to be fairer. This research helps identify a new task type for admissions tests and instructional assessment. Appendixes contain sorting task directions, the study questionnaire, and the similarity rating task directions. (Contains 4 figures, 15 tables, and 15 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it
☐ Minor changes have been made to improve reproduction quality
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
H. I. BRAUN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED 393 935

GRE®
RESEARCH

A Task Type for Measuring the Representational Component of Quantitative Proficiency

Randy Elliot Bennett
Marc M. Sebrechts
and
Donald A. Rock

August 1995

GRE Board Professional Report No. 92-05P
ETS Research Report 95-19

Educational Testing Service, Princeton, New Jersey

BEST COPY AVAILABLE

A Task Type for Measuring the Representational Component of Quantitative Proficiency

Randy Elliot Bennett
Marc M. Sebrechts
and
Donald A. Rock

GRE Board Report No. 92-05P

August 1995

This report presents the findings of a research project funded by and carried out under the auspices of the Graduate Record Examinations Board.

Educational Testing Service, Princeton, NJ 08541

****

Researchers are encouraged to express freely their professional judgment. Therefore, points of view or opinions stated in Graduate Record Examinations Board Reports do not necessarily represent official Graduate Record Examinations Board position or policy.

****

The Graduate Record Examinations Board and Educational Testing Service are dedicated to the principle of equal opportunity, and their programs, services, and employment policies are guided by that principle.

EDUCATIONAL TESTING SERVICE, ETS, the ETS logo, GRADUATE RECORD EXAMINATIONS, and GRE are registered trademarks of Educational Testing Service.

Copyright © 1995 by Educational Testing Service. All rights reserved.

BEST COPY AVAILABLE

Abstract

Two computer-based categorization tasks were developed and pilot tested. For Study I, the task asked examinees to sort mathematical word problem stems according to prototypes. Results showed that those who sorted well tended to have higher GRE General Test scores and college grades than did examinees who sorted less proficiently. Examinees generally preferred this task to multiple-choice items like those found on General Test quantitative section and felt the task was a fairer measure of their ability to succeed in graduate school. For study II, the task involved rating the similarity of item pairs. Both mathematics test developers and students participated, with the results analyzed by individual differences multidimensional scaling. Experts produced more scaleable ratings overall and primarily attended to two dimensions. Students used the same two dimensions with the addition of a third. Students who rated like the experts in terms of the dimensions used tended to have higher admissions test scores than those who used other criteria. Finally, examinees preferred multiple-choice questions to the rating task and felt that the former was a fairer indicator of their scholastic abilities. The major implication of this work is in identifying a new task type for admissions tests, as well as for instructional assessment products that might help lower scoring examinees localize and remediate problem-solving difficulties.

BEST COPY AVAILABLE

Acknowledgments

Many people contributed to this project. Jan Flaugher served as project manager, taking responsibility for coordinating tutorial development, computer-based item production, development of the response types, and local pilot testing, among other things. Daryl Ezzo, Mary Morley, and Beth Brownstein wrote the test items. They, along with Jeff Wadkins, provided invaluable criticism leading to important improvements in the tasks. Holly Knott and Helen Lambert produced the on-screen versions of the items, and Jeff Jenkins and Joel Goldberger programmed the response types. Peggy Redman conducted pilot testing, and oversaw data collection and data entry. Altamese Jackenthal and Eleanore DeYoung supplied administrative support. Thanks are also owed to Maria Potenza and the GRE graduate assistants who collected field test data, and to the undergraduates who gave their time to take the test. Finally, appreciation is expressed to the GRE Program for its administrative support and to the GRE Board for funding of this project.

A Task Type for Measuring the Representational Component of Quantitative Proficiency

Cognitive research has led to a characterization of experts as individuals who have large, highly organized domain-specific knowledge bases; who can solve problems rapidly but who appear to spend a disproportionate amount of solution time in planning; who are adept at constructing runnable mental models of a problem situation; who are skilled at using self-regulatory processes; and whose knowledge is tightly bound to conditions of use and is highly proceduralized (Glaser, 1991).

Experts also are able to perceive large, meaningful, integrated patterns in a problem situation quickly--that is, to represent the situation rapidly in terms of its underlying solution structure with only secondary attention to surface features (Glaser, 1991). In contrast, the patterns novices detect are smaller, less organized, more literal and surface oriented, and much less related to underlying principles (Chi, Feltovich, & Glaser, 1981).

Much of the expert's power lies in this rapid representational ability (Chi, Feltovich, & Glaser, 1981). Considerable evidence exists to suggest that experts represent problems by category and that these categories direct problem solving. Categories are thought to direct problem solving by eliciting a knowledge structure or schema that, at least for experts, includes potential solution methods. The initial categorization, thus, restricts search to a small range of potential solution paths.

Although the classic work on expert-novice differences in problem categorization was done in chess (Chase & Simon, 1973) and physics (Chi, Feltovich, & Glaser, 1981), a good deal of research in this area has been undertaken in mathematical problem solving. Hinsley, Hayes, and Simon (1977) conducted several experiments, one of which asked high school and college students to sort according to "problem type" 76 algebra word problems selected from a high school text. They found that subjects were able to recognize and agree upon a limited set of categories. From the follow-up experiments, the researchers concluded that recognition often occurred quickly, that subjects had information about the categories that was useful for solution, and that subjects could and often did use this information, even when the instructions simply were to solve the problems and did not call attention to problem classification.

Gliner (1989) asked college students to sort 13 word problems varying in surface detail as well as solution method according to "mathematical structure" and then to solve the questions. She found the better problem solvers sorted problems based on underlying mathematical relationships, whereas the poorer problem solvers relied on characteristics such as the cover story context and the units of measurement.

Silver (1979) designed a set of 24 word problems that could be grouped according to surface structure or solution procedure. He asked junior high school students to sort the problems into "mathematically related" groups and to solve them. Mathematical achievement and ability test scores from several standardized instruments also were collected. Silver found that the tendency to sort on the basis of mathematical structure was significantly correlated with the ability and achievement measures. In a subsequent study, Silver

BEST COPY AVAILABLE

(1981) concluded that good problem solvers tended to recall accurately the structure of the problems they had solved, whereas poor problem solvers did not.

The above studies suggest cross-sectional differences in categorization ability among problem solvers who vary in mathematical skill and general ability. Investigators have also looked at how categorization ability changes with skill development. Schoenfeld and Herrmann (1982) asked college students and mathematics professors to sort 32 mathematical word problems into groups based on which problems "were similar mathematically in that they would be solved in the same way" (p.486). Two groups of students sorted the problems. One group sorted before and after completing a 45-hour, one-month course on mathematical problem solving and the other before and after a similarly intensive course in structured programming. Before instruction, the performance of students in both groups showed the usual expert-novice difference vis-a-vis the performance of professors. However, after instruction, students taking the mathematics course perceived problem relatedness more like the experts, whereas the students taking the structured programming course did not. Similar improvements in categorization as a function of the development of domain knowledge have also been reported in other content areas (e.g., Wagner, Sebrechts, & Black, 1985).

The purpose of the current study was to develop and pilot test a prototype computer-based categorization task. What are the potential advantages of such a task for the GRE Program? First, the categorization task has a strong theoretical basis in cognitive research. As such, it represents an opportunity to incorporate a task type linked to proficiency through cognitive principles. Such links are rarely found for the item formats typically used in standardized tests.

Second, the categorization task is usable in multiple domains. It can be employed in any subject-matter area that utilizes word problems in which the solution structures can be described in terms of underlying principles or solution methods. Chemistry, physics, and mathematics are examples.

Third, the task can be performed with reasonable efficiency because examinees need only categorize problems, not solve them. For example, Schoenfeld and Herrmann (1982) reported that college students were able to sort 32 items comfortably in 20 minutes.

Finally, there is reason to believe that the responses can be objectively and automatically scored. Scoring might occur on two dimensions. The more obvious dimension is the extent to which the examinee's categorization reflects the underlying structural similarities present in the item set. For scoring purposes, this underlying structure is arrived at through design--by writing items so they are classifiable on the basis of deep structure or surface features. This hypothesized structure then becomes a tentative scoring key to be verified by the classifications of experts, the results of which are used to revise the item set as necessary. A second potential scoring dimension is response latency: Experts make categorizations very rapidly. Within levels of classification skill, additional proficiency information might be gained from the time needed to arrive at a response.

The cited research suggests that categorization performance is an index of the ability to represent problems. In the research literature, categorization tasks take two basic forms. One form presents the student with word problem stems, each one printed on an index card, and asks that the problems be sorted according to solution structure into as many piles as necessary. In an alternative formulation, subjects are asked to rate each possible pair of problems according to their structural similarity.

For this project, we created computer-based instruments loosely based on each method. The instruments were targeted primarily at applicants to graduate programs in fields other than the hard sciences and mathematics. Our exploration of the sorting task is described as Study I and that for similarity rating as Study II.

STUDY I: Sorting

Method

Subjects

Two groups of subjects were used for this pilot study. Nine faculty members constituted the first group. Three were from the engineering department at the University of Maryland and six from the physics and the economics and business departments of The Catholic University of America.

The second group consisted of 107 paid undergraduate volunteers recruited through 11 institutions of higher education in different regions of the United States, as well as through local community resources. The majority had majored in the social sciences, humanities/arts, business, or education and intended to pursue graduate education in similar fields (see Table 1). More than two thirds of the examinees were female and a third were from minority groups. Of the 101 who indicated plans to pursue graduate education, 35% viewed the Ph.D. degree as their goal.

GRE General Test scores were available for 65 of the subjects. The presence of these scores was independent of whether or not examinees were from the majority or minority groups ($r = .07$, $p > .05$), gender ($r = .04$, $p > .05$), and graduate degree goal ($r = -.04$, $p > .05$). It was significantly related to undergraduate grade-point average (UGPA), however: those with higher UGPAs tended have already taken the GRE ($r = .36$, $p < .01$).

SAT Mathematical scores were available for 54 subjects. Availability was unrelated to gender ($r = .11$, $p > .05$), minority group membership ($r = .04$, $p > .05$), or the presence of General Test scores ($r = -.03$, $p > .05$). It was related to self-reported math rating ($r = .20$, $p < .05$), UGPA ($r = .23$, $p < .05$), and degree goal ($r = .24$, $p < .05$). Those who viewed themselves as more mathematically adept, performed better in college, and intended to pursue the Ph.D. were more likely to have reported SAT Mathematical scores.

Instruments

Sorting task. Figure 1 gives the primary screen for the sorting task. To facilitate scoring, the task was changed from creating an undetermined number of categories to matching the target problem with one of four

BEST COPY AVAILABLE

Table 1
Demographic Data

Background Characteristic	Number Responding	Sample Value
Percentage Female	107	69%
Percentage Non-White	106	36%
Percentage with Ph.D. Goal	99	35%
Undergraduate Major	107	---
Social Sciences	---	36%
Life Sciences	---	5%
Physical Sciences	---	1%
Humanities/Arts	---	27%
Business	---	14%
Engineering	---	2%
Education	---	8%
Other	---	8%
Intended Graduate Major	101	---
Social Sciences	---	34%
Life Sciences	---	6%
Physical Sciences	---	1%
Humanities/Arts	---	20%
Business	---	15%
Engineering	---	0%
Education	---	13%
Other	---	12%
GRE Quantitative Mean (SD)	65	546 (118)
GRE Verbal Mean (SD)	65	502 (109)
GRE Analytical Mean (SD)	65	576 (115)
SAT Mathematical Mean (SD)	54	564 (93)
UGPA Mean (SD)	107	3.1 (.5)
Math Proficiency Mean (SD)	107	2.5 (.7)

Note. SAT and UGPA are self-reported, with UGPA on a 1-4 scale. Math proficiency is a rating on a 1-4 scale of the examinee's perception of his or her own mathematical skill, with 4 indicating the highest proficiency level.

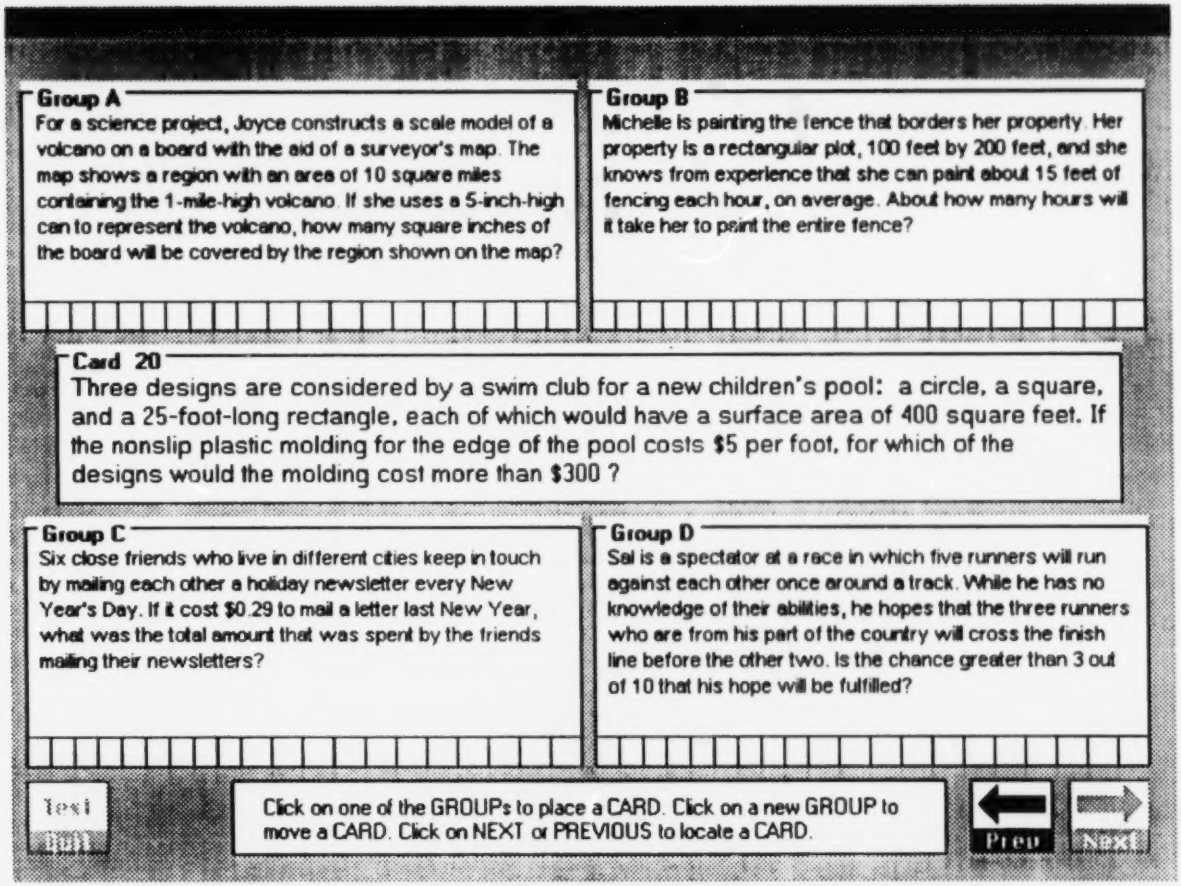

Group A
For a science project, Joyce constructs a scale model of a volcano on a board with the aid of a surveyor's map. The map shows a region with an area of 10 square miles containing the 1-mile-high volcano. If she uses a 5-inch-high can to represent the volcano, how many square inches of the board will be covered by the region shown on the map?
Group B
Michelle is painting the fence that borders her property. Her property is a rectangular plot, 100 feet by 200 feet, and she knows from experience that she can paint about 15 feet of fencing each hour, on average. About how many hours will it take her to paint the entire fence?
Card 20
Three designs are considered by a swim club for a new children's pool: a circle, a square, and a 25-foot-long rectangle, each of which would have a surface area of 400 square feet. If the nonslip plastic molding for the edge of the pool costs $5 per foot, for which of the designs would the molding cost more than $300 ?
Group C
Six close friends who live in different cities keep in touch by mailing each other a holiday newsletter every New Year's Day. If it cost $0.29 to mail a letter last New Year, what was the total amount that was spent by the friends mailing their newsletters?
Group D
Sal is a spectator at a race in which five runners will run against each other once around a track. While he has no knowledge of their abilities, he hopes that the three runners who are from his part of the country will cross the finish line before the other two. Is the chance greater than 3 out of 10 that his hope will be fulfilled?
Test
Quit
Click on one of the GROUPs to place a CARD. Click on a new GROUP to move a CARD. Click on NEXT or PREVIOUS to locate a CARD.
Prev
Next

Figure 1. Sorting task interface.

BEST COPY AVAILABLE

prototypes. Target problems are presented in the center window; the prototypes are given in the windows above and below the target. Each prototype represents a group. To assign a target problem, the examinee clicks on a prototype, causing the target problem's number to appear in the array below the selected group.

Because the test developers working on the sorting task had concerns about whether it could be objectively scored, a second version was created that included a canonical equation with each prototype, thereby making the keys more defensible (see Figure 2).

Whether equations are included, this format is akin to multiple choice. The major difference is that in conventional tests, questions tend to be unrelated, assessing relatively isolated bits of knowledge. In problem sorting, items are written to a coherent test-level organization that the examinee reconstructs by virtue of making the correct categorizations. Thus, the examinee must consider problem relatedness in addition to bringing to bear the specific knowledge needed to understand any given item.

Tables 2 and 3 give the test-level organizations for the two 20-item sorting tasks. For the first task (Sorting I), the organization was two dimensional and involved crossing item-level mathematical structure with surface content. The former dimension was "chained" in that each group of items was related to its adjacent group(s) by sharing part of that group's mathematical structure (see Table 2). So, for example, combination/rate problems shared structure with rate/perimeter problems in that the solutions to problems in both sets involved rates. Structure was considered to include such features as the principle(s) required for solution, the form of the equation(s) that model the problem, the methods that could lead to a solution, the role of variables, and the nature of the steps or operations. The organization for the second task (Sorting II) was simpler, using only the structural dimension, which was composed of four unrelated categories defined symbolically. For this task, the definition of "structure" was sharpened: two problems were considered to be structurally identical if every mathematical model that could be used to solve one could be used to solve the other.

Both sorting tasks were timed in two sections: eight minutes to review the four prototypes (without access to the 20 problems) and 24 minutes to sort the problems. These limits were selected so as to encourage students to sort the problems without working the solutions completely.

For the first task, five scores were derived. The first score was a count of the number of correct categorizations. The second was the number of instances in which the item was matched to its surface category but not its structural category. Third, a partial-credit score was generated in which two points were awarded if the item was matched to the correct structural category and one point if the item was matched to an adjacent category sharing some aspect of mathematical structure. The fourth and fifth scores were (a) the time from presentation of the first item to the first answer given to the 20th item (partial time) and (b) the time from presentation of the first item to quitting the test (total time). For the second sorting task, only the number of correct categorizations and the two timing scores were used.

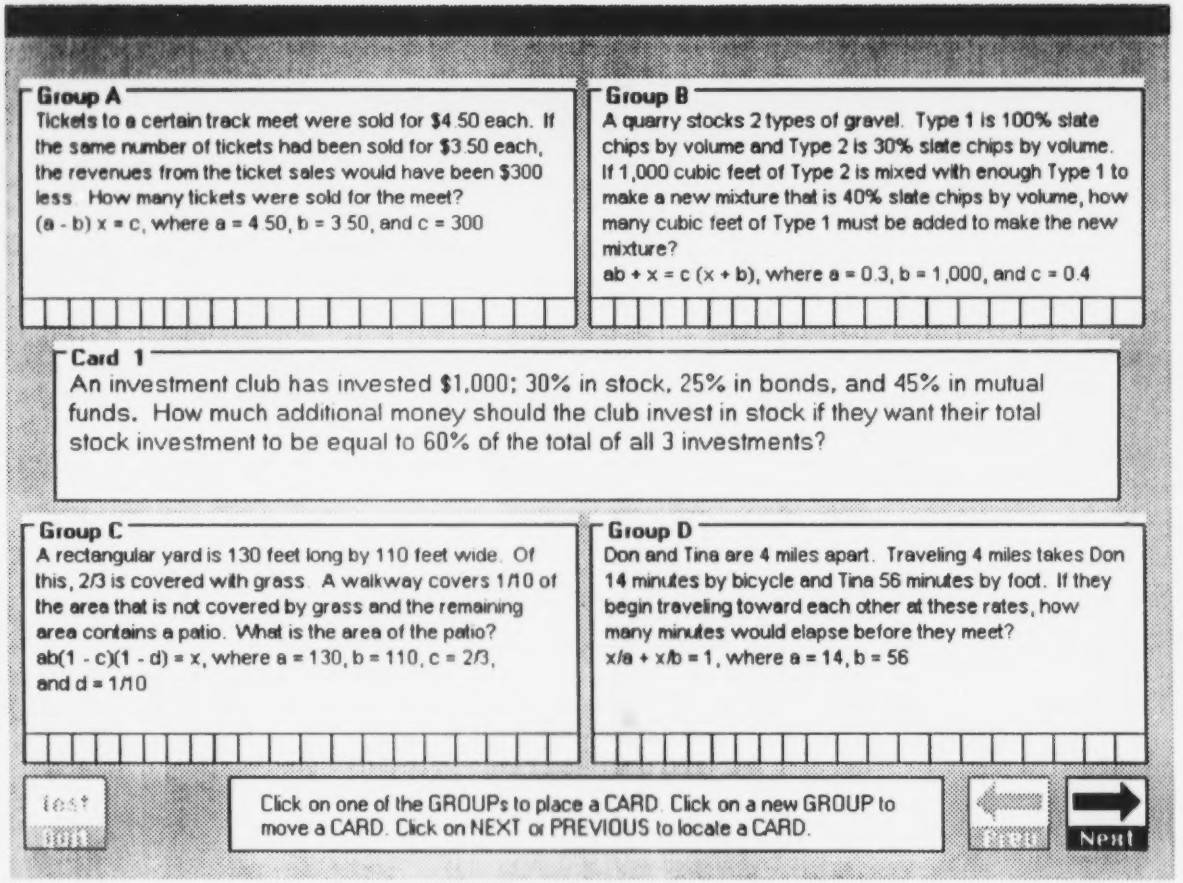

Group A
Tickets to a certain track meet were sold for $4.50 each. If the same number of tickets had been sold for $3.50 each, the revenues from the ticket sales would have been $300 less. How many tickets were sold for the meet?
(a - b) x = c, where a = 4.50, b = 3.50, and c = 300
Group B
A quarry stocks 2 types of gravel. Type 1 is 100% slate chips by volume and Type 2 is 30% slate chips by volume. If 1,000 cubic feet of Type 2 is mixed with enough Type 1 to make a new mixture that is 40% slate chips by volume, how many cubic feet of Type 1 must be added to make the new mixture?
ab + x = c (x + b), where a = 0.3, b = 1,000, and c = 0.4
Card 1
An investment club has invested $1,000; 30% in stock, 25% in bonds, and 45% in mutual funds. How much additional money should the club invest in stock if they want their total stock investment to be equal to 60% of the total of all 3 investments?
Group C
A rectangular yard is 130 feet long by 110 feet wide. Of this, 2/3 is covered with grass. A walkway covers 1/10 of the area that is not covered by grass and the remaining area contains a patio. What is the area of the patio?
ab(1 - c)(1 - d) = x, where a = 130, b = 110, c = 2/3, and d = 1/10
Group D
Don and Tina are 4 miles apart. Traveling 4 miles takes Don 14 minutes by bicycle and Tina 56 minutes by foot. If they begin traveling toward each other at these rates, how many minutes would elapse before they meet?
x/a + x/b = 1, where a = 14, b = 56
Click on one of the GROUPs to place a CARD. Click on a new GROUP to move a CARD. Click on NEXT or PREVIOUS to locate a CARD.
Next

Figure 2. Sorting task with canonical equations for each prototype.

BEST COPY AVAILABLE

Table 2
Test-Level Structure for Sorting Task I

Mathematical Structure	Surface Content: Racing/ Track	Friends/ Club	Fence/ Yard	School Project
Area/Ratio	10,16	13	15	A,5
Rate/Perimeter	6	14,20	B,2	4,17
Combination/Rate	1	C,7	9,11	18
Probability/Combination	D	19,12	8	3

Note. Letter designations indicate prototype problems. Numerical designations indicate problems to be matched.

Table 3
Test-Level Structure for Sorting Task II

Mathematical Structure	Item Number
$(a - b)x = c$	2, 7, 14, 17, 18, A
$ab + x = c(x + b)$	1, 9, 13, 15, 16, B
$ab(1-c)(1-d) = x$	3, 4, 8, 11, 19, C
$x/a + x/b = 1$	5, 6, 10, 12, 20, D

Note. Letter designations indicate prototype problems. Numerical designations indicate problems to be matched.

Opinion and background questionnaire. Each examinee was asked to complete a brief questionnaire asking for background information and for perceptions of the task, its timing, difficulty, and computer-based presentation.

Procedure

Faculty members were asked to complete paper-and-pencil versions of both sorting tasks as a means of verifying their keys. Faculty responses were then compared with the developers' keys. In cases where a faculty response differed from the key, the faculty member was sent a form. The form contained the person's proposed answer along with the alternative, and asked the faculty member to indicate whether the alternative was a better choice after reexamination, was equally plausible, was acceptable as a second choice, or was not acceptable at all. In addition, the form requested a description of the reason for giving the proposed answer and, where applicable, for rejecting the alternative.

For students, a short information booklet that explained the task was given out upon recruitment. At the test session, examinees were given more detailed instructions, including directions for using the computer interface.

All 107 examinees took the first sorting task. Sixty of these individuals also took the second task, with the two tasks administered in counterbalanced order. Appendix A shows the test directions for the second sorting task.

Data Analysis

For the faculty responses, the number of disagreements with the key was tallied before and after the requests for confirmation. Reasons for changing or retaining the original answer were categorized.

The student data were analyzed to provide information about the meaning of sorting scores and the students' perceptions of the task. With respect to meaning, reliability was examined, as well as relevance to mathematical reasoning skill and to college performance more generally. Reliability was estimated by computing coefficient alpha for the number-right scores on each of the two sorting tasks. Relevance was addressed in two ways. First, the product-moment correlations were examined between sorting scores and admissions scores, self-reported math rating, and undergraduate grade-point average. Second, least-squares linear multiple regression was used to determine how the General Test scales incrementally added to the explanation of sorting scores--thereby giving a better sense of the skills required for success on the task. Finally, examinee responses to the questionnaire were tabulated.

Results

On the first 20-item sorting task, all faculty agreed with the key for 9 of the items; eight of the nine faculty agreed on 16 of the items. After reconsideration, unanimous agreement was found for 14 items and agreement among eight of nine faculty for 18 questions. For the remaining two questions, more than one judge felt multiple prototypes were keyable.

Discussions with the faculty members suggested that, for future studies, several of the problems might be revised to reduce ambiguities that encouraged matching with prototypes in addition to, or other than, those intended.

The second 20-item sorting task was composed of a new set of problems accompanied by equations; thus, the results may differ because of changes in the items and prototypes, clearer definition of the concept of structure, or including the equations. For this task, there was unanimous agreement on the keys for 15 items and agreement among eight of nine judges on 19 items. For the five items with imperfect agreement, discussion with the individuals showed the disagreements to be caused by superficial mistakes on the judges' part. After reconsideration, all judges agreed on all items.

Table 4 gives summary statistics for students' performance on the sorting tasks and associated measures. The means and standard deviations appear similar for the score and timing indicators on the two sorting tasks. Of note is that the distribution of partial-credit scores and of surface scores is considerably skewed, with most subjects having relatively high partial-credit scores and very low surface scores.

Coefficient-alpha reliabilities for the first and second sorting tasks' number-right scores were .51 and .68, respectively. (Each score represents about half an hour of testing time.) Doubling the number of problems would bring the first task's reliability to .68 and the second task's internal consistency to .81. Tripling the length would produce values of .76 and .87, respectively. These values are below those that would be expected from conventional multiple-choice tests with comparable administration times.

Table 5 presents the correlations among the Sorting I scores and the criterion variables. Number-right score was significantly related to the mathematical reasoning admissions measures (r with GRE-Q = .53, $p < .01$; r with SAT-M = .52, $p < .01$;), as well as to the verbal and analytical reasoning tests, for which correlations ran in the mid-forties. This score was also strongly negatively related to surface score ($r = -.64$, $p < .01$), an expected result given that those who sorted well would necessarily have low surface scores and those who sorted poorly might have higher ones by chance alone. Finally, number-right score was significantly related to college performance (r with UGPA = .40, $p < .01$).

Besides number-right score, this sorting task generated four other measures. Partial-credit score was almost perfectly correlated with the number-right score and had essentially the same pattern of relations with other variables. Surface score was significantly related to all admissions test scores and to UGPA, its values extending from the low thirties to low forties. These relations were negative, consistent with this score's purpose as a measure of how frequently the student mistook surface features for deep structure. Finally, the partial-time index was significantly related to GRE-V ($r = -.31$, $p < .05$) and GRE-A ($r = -.26$, $p < .05$). The relation of partial time to GRE-Q was not significant, nor were the relations between total time and the General Test scales. However, all six relations between the time measures and General Test scores were negative. In this sample, better verbal, analytical, and mathematical reasoners sorted more quickly--in addition to doing so more accurately--than less adept students.

BEST COPY AVAILABLE

Table 4
Summary Statistics for Students' Performance

Variable	Scale	N	Mean	SD	Skewness
Sorting Task I					
Number-Right Score	0-20	107	14.9	2.5	-.5
Partial-Credit Score	0-40	107	33.2	3.6	-1.1
Surface Score	0-17	107	1.3	1.4	2.1
Partial Time (sec)	0-1440	107	886	246	.2
Total Time (sec)	0-1440	107	1103	285	-.4
Sorting Task II					
Number-Right Score	0-20	60	14.6	2.8	.1
Partial Time (sec)	0-1440	60	919	283	-.1
Total Time (sec)	0-1440	60	1157	292	-.8
GRE Quantitative	200-800	65	546	118	.2
GRE Verbal	200-800	65	502	109	.5
GRE Analytical	200-800	65	576	115	-.3
SAT Mathematical	200-800	54	564	93	.6
UGPA	1-4.0	107	3.1	.5	-.4
Math Proficiency Rating	1-4	107	2.5	.7	.0

Table 5

Correlations Among Students' Sorting I Scores and Criterion Variables (n=107)

	Number-Right Score	P-C Score	Surface Score	Part. Time	Total Time	UGPA	Math Rating	GRE-Q	GRE-V	GRE-A
Sorting I Scores										
Part-Credit Score	.94**									
Surface Score	-.64**	-.62**								
Partial Time	-.01	.02	-.07							
Total Time	.03	.05	-.08	.71[a]						
Criterion Variables										
UGPA	.40**	.38**	-.43**	.01	-.02					
Math Rating	.18	.15	.05	-.13	-.18	.13				
GRE Quantitative	.53**	.53**	-.38**	-.15	-.19	.45**	.50**			
GRE Verbal	.47**	.44**	-.33**	-.31*	-.17	.41**	.04	.57**		
GRE Analytical	.44**	.39**	-.31*	-.26*	-.20	.34**	.26	.66**	.70**	
SAT Mathematical	.52**	.53**	-.30*	.04	-.13	.33*	.46*	.88**	.45**	.51**

Note. For SAT-M, n = 54 and for GRE n = 65, except for correlation of SAT-M with GRE for which n = 32.

[a]Part-whole correlation.

* $p < .05$

** $p < .01$

In Table 6 are the correlations of scores on the second sorting task with those from the first sorting task and with the criterion variables. In interpreting these scores, it is important to note that these data are for a subset of the sample that took the first sorting task. In this subsample, the two sorting tasks had a correlation of .46 ($p < .01$), about 13 points below the (geometric) mean reliability of the two scales, suggesting that they might not be measures of the same construct. The sorting II number-right score was related to the mathematical reasoning tests (GRE-Q $r = .66$, $p < .01$; SAT-M $r = .60$, $p < .01$), as well as to self-reported math rating ($r = .47$, $p < .01$). The correlations with the mathematical reasoning tests ran from 13 to 19 points below the measures' average reliabilities. The relations with verbal and analytical reasoning were also significant, but in the forties, hinting that the reasoning evoked was more mathematical than verbal or analytical. As for the first task, this sorting task also was significantly related to UGPA ($r = .35$, $p < .01$). Finally, all six relations with time were again negative, although only one, total time, was significantly related to General Test performance (GRE-A $r = -.30$, $p < .05$).

To get a clearer sense of the role of different reasoning skills in sorting performance, we regressed number-right and surface sorting scores separately on the three General Test scales in turn. Because of the small sample sizes, the results of these analyses should be considered suggestive only. GRE-Q was entered first, followed by GRE-V and GRE-A. This order followed the presumption that the task primarily reflected mathematical reasoning skills but that verbal and analytical skills beyond those tapped by GRE-Q might also be required (e.g., verbal skills to understand the word problem and analytical ones to help map relations between problems). As Table 7 shows, the results for the first sorting task appear to bear out the expectation in that GRE-V makes a significant incremental contribution to explaining number-right score; surface score, however, appears to be largely a function of mathematical reasoning or of more general reasoning skills shared by the three scales. For the second task, GRE-Q was the only significant predictor of number-right score (this task did not generate a surface score); its partial regression weight was four times that of GRE-V and many more times that of GRE-A. This result could be due to the addition of equations for the prototypes in the second task or to the sharpening of the definition of structure, either of which might reduce the need to reason verbally.

Table 8 summarizes the subjects' questionnaire responses. As can be seen, the response distributions were very similar for questions asked about each task. Roughly 90% of examinees found the test to be right in difficulty and an even greater percentage thought the timing was at least adequate. Most examinees indicated they would prefer to take the sorting task over the kinds of multiple-choice questions found on the General Test and that the sorting task was a fairer representation of their ability to undertake graduate study. Subjects preferred taking a computer-based test to paper-and-pencil about 2 to 1 (51% to 28%). With respect to computer use, over 90% indicated using a computer weekly or more, and almost that percentage said they used a computer to write papers for school. Finally, 91% responded that the computer was easy to use for the sorting task; only 9% percent found using it somewhat difficult.

BEST COPY AVAILABLE

Table 6
Correlations of Sorting II Scores with Sorting I Scores and Criterion Variables (n=60)

	Sorting II		
	Number-Right Score	Partial Time	Total Time
Sorting I			
Number-Right Score	.46**	-.01	.09
Partial-Credit Score	.42**	.00	.14
Surface Score	-.31*	-.05	-.10
Partial Time	-.07	.59**	.54**
Total Time	.01	.48**	.65**
UGPA	.35**	.00	.19
Math Rating	.47**	-.16	-.06
GRE Quantitative	.66**	-.04	-.03
GRE Verbal	.47**	-.09	-.11
GRE Analytical	.48**	-.18	-.30*
SAT Mathematical	.60**	-.04	-.18

Note. SAT Mathematical n = 29 and GRE n = 55.

* $p < .05$
** $p < .01$

Table 7
Multiple Regression of Sorting I and II Scores on Students' GRE General Test Scores

Independent Variable	R	R^2	Increment in R^2	Incremental F	p	Standardized Regression Weight
Sorting I Number-Right Score (n=65)						
1. GRE Quantitative	.53	.28	.28	24.5	.00	.38*
2. GRE Verbal	.57	.32	.04	4.1	.05[a]	.25
3. GRE Analytical	.57	.32	.00	.0	.92	.02
Sorting II Number-Right Score (n=55)						
1. GRE Quantitative	.66	.44	.44	41.1	.00	.61**
2. GRE Verbal	.67	.45	.01	.9	.35	.14
3. GRE Analytical	.67	.45	.00	.1	.78	-.05
Sorting I Surface Score (n=65)						
1. GRE Quantitative	.38	.15	.15	10.9	.00	-.30
2. GRE Verbal	.41	.17	.02	1.3	.27	-.16
3. GRE Analytical	.41	.17	.00	.0	.99	.00

[a]Value rounded from .048.
* $p < .05$
** $p < .01$

Table 8
Students' Questionnaire Responses

Question	Sorting Task I	Sorting Task II
How easy was the test?		
Too easy	6%	0%
About right	89%	92%
Too difficult	6%	8%
How adequate was the timing?		
Too little	3%	7%
About right	74%	67%
Too much	23%	27%
Which kind of question would you rather take?		
Regular multiple-choice	30%	35%
Problem representation	54%	52%
No preference	16%	13%
Which kind of question is a fairer indicator?		
Regular multiple-choice	23%	23%
Problem representation	59%	67%
No preference	19%	10%
Which kind of test would you rather take?		
Paper-and-pencil	28%	---
Computer-based	51%	---
No preference	22%	---
How often have you used a computer this year?		
Never or almost never	7%	---
About once a week	45%	---
Daily or almost daily	49%	---
How do you usually write a paper for school?		
Pencil (or pen) and paper	10%	---
Typewriter	1%	---
Computer	89%	---
How easy was the computer to use for the test?		
Very easy	91%	---
Somewhat difficult	9%	---
Very difficult	0%	---

Note. Questions were edited for tabular presentation. The complete questionnaire is given in Appendix B. N = 107 for Sorting Task I and 60 for Sorting Task II (all of whom took Sorting Task I).

STUDY II: Similarity Rating

Method

Subjects

Two groups of subjects were used for this pilot study. Five ETS mathematics test developers constituted the first group. The second group consisted of 35 paid undergraduate volunteers from 11 institutions located in different regions of the continental United States (see Table 9). Most members of the latter group had majored in the social sciences, humanities/arts, or education and intended to pursue graduate work in a similar field.

Instruments

Similarity rating task. Figure 3 gives the interface for the similarity rating task, which requires the examinee to rate all possible pairs in a pool of nine problems. To make a rating, the examinee clicks the mouse on a cell in the matrix. This action causes a problem to appear in the upper left window corresponding to the row number of the highlighted cell and a problem to appear in the upper right window related to the cell's column designation. The examinee enters a similarity rating by clicking on the scale to the right of the matrix, where "1" indicates that the problems are "very different" in underlying structure (relative to other items in the set) and "9" means that they are the same (relative to the other items).

The nine-item set was similar to the first sorting task in its underlying design and also shared some items with it (see Table 10). As for the sorting tasks, timing was in two sections: Five minutes was given to review the nine items as a set and 30 minutes to rate the 36 possible pairs. Two scores were generated: a measure of the extent to which the examinee used the same dimensions as test developers and the time spent from presentation of the rating matrix to quitting the test.

Opinion and background questionnaire. Each examinee was asked to complete a questionnaire similar to the one used for examinees taking the sorting tasks.

Procedure

Screen prints describing the task and giving directions for using the computer interface were given out upon recruitment. At the test session, examinees reviewed this information on screen before taking the test. Appendix C shows the test directions for the rating task.

BEST COPY AVAILABLE

Table 9
Demographic Data

Background Characteristic	Number Responding	Sample Value
Percentage Female	35	57%
Percentage Non-White	34	44%
Percentage with Ph.D. Goal	34	41%
Undergraduate Major	35	---
Social Sciences	---	43%
Life Sciences	---	14%
Physical Sciences	---	3%
Humanities/Arts	---	14%
Business	---	3%
Engineering	---	0%
Education	---	6%
Other	---	17%
Intended Graduate Major	35	---
Social Sciences	---	37%
Life Sciences	---	17%
Physical Sciences	---	6%
Humanities/Arts	---	11%
Business	---	3%
Engineering	---	0%
Education	---	9%
Other	---	17%
GRE-Q Mean (SD)	20	640 (83)
SAT-M Mean (SD)	23	603 (91)

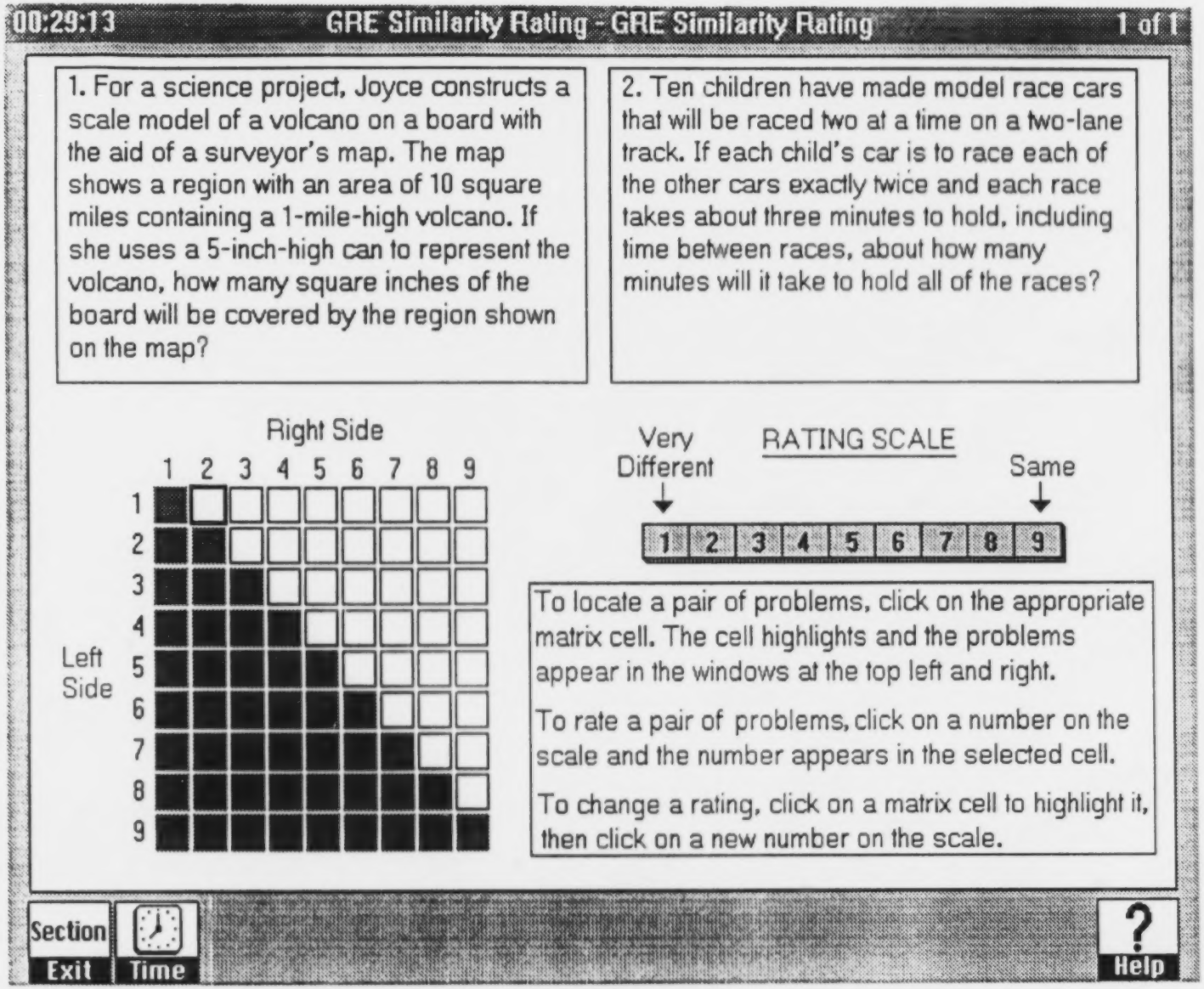

00:29:13
GRE Similarity Rating - GRE Similarity Rating
1 of 1
1. For a science project, Joyce constructs a scale model of a volcano on a board with the aid of a surveyor's map. The map shows a region with an area of 10 square miles containing a 1-mile-high volcano. If she uses a 5-inch-high can to represent the volcano, how many square inches of the board will be covered by the region shown on the map?
2. Ten children have made model race cars that will be raced two at a time on a two-lane track. If each child's car is to race each of the other cars exactly twice and each race takes about three minutes to hold, including time between races, about how many minutes will it take to hold all of the races?
Right Side
1 2 3 4 5 6 7 8 9
Left Side
Very Different
RATING SCALE
Same
1 2 3 4 5 6 7 8 9
To locate a pair of problems, click on the appropriate matrix cell. The cell highlights and the problems appear in the windows at the top left and right.
To rate a pair of problems, click on a number on the scale and the number appears in the selected cell.
To change a rating, click on a matrix cell to highlight it, then click on a new number on the scale.
Section
Exit
Time
Help

Figure 3. Interface for the similarity rating task.

BEST COPY AVAILABLE

Table 10
Test-Level Structure for the Similarity Rating Task

Mathematical Structure	Surface Content		
	Racing/ Track	Fence/ Yard	School Project
Area/Ratio	5	7	1
Rate/Perimeter	8	3	4
Combination/Rate	2	6	9

Note. Numbers in table are item designations.

Data Analysis

Expert (i.e., test developer) and examinee ratings were analyzed by individual differences multidimensional scaling as implemented in the ALSCAL program (Young & Lewyckyj, 1979). The individual difference multidimensional scaling model attempts to map psychological distances onto a group stimulus space whose underlying dimensions are assumed to reflect the structure of the stimuli (in this case, items) as perceived by the entire group of subjects. The individual differences model also yields a weight matrix in addition to the coordinates of the stimuli on the dimensions. The weight matrix quantifies how salient each dimension was in determining each individual's judgments about the similarity between pairs of items. An overall goodness of fit measure (i.e., R^2) for each subject indicates the proportion of variation in an individual's judgments that are explained by the group model. The relative extent to which an individual used a particular group dimension can be estimated by dividing the squared subject weight on that dimension by the sum of the individual's squared weights.

ALSCAL analyses of the experts' data were used to model their judgments and determine how well that model fit the intended underlying structure of the item set. The resulting model was compared with an ALSCAL model derived from students' ratings to see if the two groups scaled items on the same basis and, if not, how they differed. For each student, the proportions of variance attributable to the dimensions that were and were not used by experts were computed as proficiency indices and descriptively related to external criteria.

Results

Two individual differences multidimensional scaling solutions were computed to explain the test developers' similarity judgments for the nine items. The first solution was three-dimensional and the second two-dimensional. In keeping with directions to the examinees, both solutions were ordinal, so that only the rankings of item pairs were considered and not their distances from one another. For the three-dimensional model, the average expert weights on the dimensions were .57, .34, and .07, respectively, suggesting that the first two dimensions were used far more heavily than the last one was. The overall R^2 values for the two solutions support this conclusion, showing only a minimal loss in fit (.99 to .95) from the three- to the two-dimensional solutions (see Table 11). Examination of the R^2 values for each test developer showed the two-dimensional model to function nearly as effectively for all but one rater (#4).

Figure 4 gives the coordinates in two-dimensional space of each of the nine items. As can be seen, three distinct groups appear. The groups are the same three mathematical structure categories used to generate the items (see Table 10). Thus, the expert composite recapitulated the intended underlying structure of the item set, lending empirical support to the composite as a scoring key. From an examination of the items and their positions in this space, the first dimension might be related to the salience of multiplying by a rate and the second to the extent the item involved geometry.

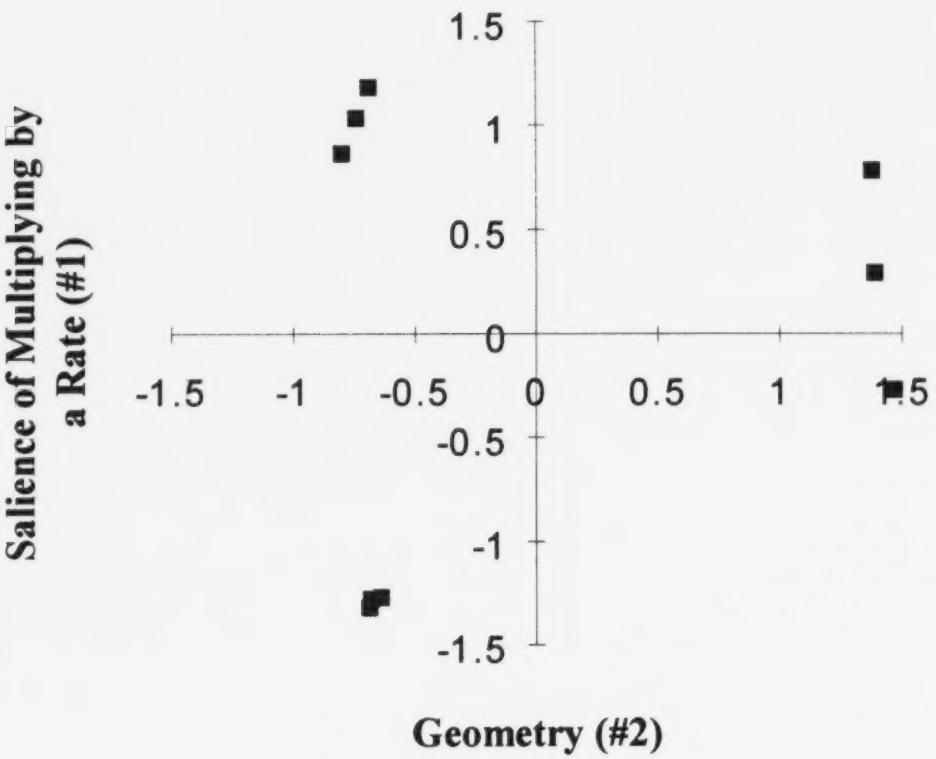
BEST COPY AVAILABLE

Table 11
Proportions of Variance Accounted for in Test Developers' Similarity Ratings by Two Different Multidimensional Scaling Solutions

	Solutions	
Test Developer	Three-Dimensional	Two-Dimensional
1	.96	.93
2	.99	.99
3	1.00	1.00
4	.98	.84
5	1.00	.99
Average	.99	.95

Figure 4

Placement of Items in Two Dimensional Space



The weights in Table 12 indicate the importance of each of the two dimensions for each developer in determining the similarity ratings. Clearly, there were differences in the extent to which individual developers used the dimensions in making their judgments.

Two- and three-dimensional models were also fit to the student data. The three-dimensional model produced an average $\underline{R}^2$ of .66 and the two-dimensional one an $\underline{R}^2$ of .59. For the three-dimensional model, the $\underline{R}^2$ values for individuals ranged from .30 to .98. This level of fit indicates that, on the whole, students' ratings were less internally consistent--or scaleable--than developers' ratings were. As might be expected, many students do not rate items using the same dimensions as experts do.

Table 13 shows the correlations between the nine stimulus weights that compose each of the three dimensions derived from the five developers and those composing the three dimensions derived from the 35 students. As the table shows, the first two dimensions are common to the experts and students: The experts' weights for dimension #1 correlate .89 with the weights used by students for their first dimension; for the second dimension, the weights correlate .95. The third dimension used by developers is clearly different from the third dimension used by students, as indicated by the .48 correlation between them. Finally, the average subject weights for the three dimensions indicate that the students used the dimensions about equally overall (.27, .20, .19), whereas the experts relied more heavily on the first two dimensions.

Further support for the similarity of the first two dimensions across experts and students was gained by combining the two samples and fitting a two-dimensional model. The proportions of variance explained for each test developer (.89, .88, .96, .80, and .94, respectively) were only slightly lower than those for the experts-only model.

One method of scoring each examinee's ratings is to calculate the proportion of scaleable variance due to using dimensions that the experts did and did not use. These values were computed by dividing each student's squared subject weights on each of the three dimensions by the proportion of variance accounted for by the model for that student (where the model was determined from a combined sample of experts and students). Table 14 gives correlations between the extent that students used each dimension in making similarity judgments and mathematical admissions test scores. Because of the extremely small sample sizes these values are of descriptive interest only, so they should not be generalized. Of note is that examinees who used dimension #2 (which the experts also used) tended to have higher admissions test scores than those who did not use this dimension. Those who used dimension #3 (which the experts did <u>not</u> use) tended to have lower test scores.

In Table 15 are examinees' questionnaire responses. Most examinees found the task's difficulty and timing about right (77% and 69%, respectively), and most (83%) found the computer very easy to use to take the test. However, most (74%) indicated that they would prefer taking a multiple-choice test like GRE-Q and that such a test would be a fairer indicator of their ability to succeed in graduate school (54%). The overwhelming majority were computer literate: 94% used computers about once a week or more and the same

Table 12
Weights for Each Test Developer on Each Dimension

	Dimension	
Test Developer	#1	#2
1	.68	.68
2	.99	.11
3	.80	.60
4	.48	.78
5	.94	.34
Average	.64	.31

Note. The average test developer weight is the mean of the squared individual weights.

Table 13
Correlations Among Developers' and Students' Dimensions

Experts'	Students' Dimension		
Dimension	#1	#2	#3
#1	.89	-.02	.47
#2	.30	.95	.03
#3	-.60	.16	.48

Table 14
Correlations Between Students' Dimensional Salience Scores and Criterion Variables

Criterion	Scaleable Variance Due to Dimension		
Variable	#1	#2	#3
GRE Quantitative	.18	.27	-.40
SAT Mathematical	-.02	.44	-.38
SAT Verbal	.07	.43	-.48
Time	-.01	.07	.11

Note. GRE Quantitative n = 20, SAT Verbal n = 19, SAT Mathematical n = 23.

BEST COPY AVAILABLE

Table 15
Students' Questionnaire Responses (n=35)

Question	Percentage Choosing Response
How easy was the test?	
Too easy	3%
About right	77%
Too difficult	20%
How adequate was the timing?	
Too little	17%
About right	69%
Too much	14%
Which kind of question would you rather take?	
Regular multiple-choice	74%
Problem representation	14%
No preference	11%
Which kind of question is a fairer indicator?	
Regular multiple-choice	54%
Problem representation	34%
No preference	11%
Which kind of test would you rather take?	
Paper-and-pencil	43%
Computer-based	31%
No preference	26%
How often have you used a computer this year?	
Never or almost never	6%
About once a week	29%
Daily or almost daily	66%
How do you usually write a paper for school?	
Pencil (or pen) and paper	0%
Typewriter	6%
Computer	94%
How easy was the computer to use for the test?	
Very easy	83%
Somewhat difficult	17%
Very difficult	0%

percentage typically used the computer to write papers for school. Even so, more said they would rather take a paper-and-pencil test than a computer-based one (43% to 31%).

Discussion

This study explored two approaches to measuring representational skill--the ability to perceive a problem's underlying structure rapidly without necessarily solving it. Research in several content domains has found proficient performers to be adept at problem representation, which they do in terms of categories. These categories, in turn, help direct problem solving (Chi, Feltovich, & Glaser, 1981).

The first approach required examinees to sort mathematical word problem stems into groups demarcated by prototypes. Those who did well on this task tended to be better reasoners (as indicated by admissions test scores) and more successful in college (as measured by UGPA) than those who sorted less proficiently. In addition, more-adept analytical reasoners completed the task faster than those who scored lower on GRE-A. Two versions of this approach were tried, one that paired formulas with the prototypes and one that presented only the problem stem prototypes themselves. The version that included formulas was more easily keyed by content experts, had higher internal consistency, and was more dominated by mathematical reasoning versus skills from other domains than the version without formulas. Because the items were not common across these two versions, the presence of formulas could not be established definitively as the cause of these differences.

In the second approach to measuring representational skill, a new sample of examinees and five test developers were asked to rate the similarity of all possible pairs of problem stems. In these samples, experts and students differed in their ratings. Experts produced more scaleable ratings overall and primarily attended to two dimensions. Students used the same two dimensions with the addition of a third. Finally, students who rated more like experts tended to have higher admissions test scores.

Examinees were also asked for their perceptions of the tasks and about their computer experience. For both the sorting and rating approaches, examinees tended to find the task difficulty and timing to be about right, and the computer easy to use to take the test. In addition, the overwhelming majority were computer literate. Dramatic differences were apparent, however, in examinees' reactions to the two approaches. Those who took the sorting task generally preferred it to multiple-choice questions and felt it was fairer than these questions in assessing the ability to succeed in graduate school. In contrast, those who took the rating task preferred the multiple-choice questions and felt they would be a fairer measure. Comments by examinees suggest that negative reactions to the rating task stemmed from two sources. One source was related to the rating scale. Examinees appeared generally comfortable with applying the two end-points (1 and 9), but were very uncomfortable with the absence of definitions for the intervening values that, by design, indicated only relative differences. Second, and perhaps more importantly, they found the task to be tedious, especially the need to compare each new rating to previous ones to produce an internally consistent set of values.

BEST COPY AVAILABLE

In considering these results, one should keep in mind several caveats. First, the study was intended as a development effort with pilot testing limited to small, unrepresentative samples. Thus, the empirical results are suggestive only, requiring replication with larger, more representative groups. Second, the criterion indicators were limited to a few conventional measures, several of which had substantial missing data.

What are the implications of this study for the GRE Program? The primary outcome of this project was a pair of cognitively based, computer-delivered response types for measuring one component of mathematical problem-solving proficiency. Although both types generally behaved in keeping with the underlying cognitive theory, the sorting task seemed more promising: It was more straightforward to score and more acceptable to examinees. The sorting task was not, however, without limitations. This item type shares some of the potentially negative characteristics of multiple choice in that it is an indirect indicator of expertise and presents options from which to choose, giving it a relatively high probability of a correct response from guessing. Also test developers had concerns about it, including that even with an equation in the stem some mathematically inventive examinees might discover unintended structural similarities leading them away from the keyed response. This is a valid concern that future empirical work could resolve.

Assuming these limitations can be addressed, the sorting task might be used in at least two ways. One use would be as another means of measuring mathematical reasoning within the GRE Revised Quantitative Reasoning measure. In this test, sorting tasks might be presented in small sets (e.g., four prototypes followed by four problems to be matched) or as individual items (e.g., four prototypes followed by one problem to be matched). The former strategy should limit, and the latter remove, the potential for unwanted dependencies in which misrepresenting one prototype causes misclassification of all problems intended to be matched with it. Reducing or removing these dependencies might raise reliability by decreasing variation in performance across problems and make the task type more amenable to adaptive testing. In considering this use, it might be worth studying whether the task would change if the prototypes consisted only of equations. This configuration would allow more prototypes to be presented for each item, lowering the probability of guessing correctly. A concern with this approach, however, is that during informal debriefing some high-performing examinees indicated ignoring the equations because they were not as confident working with mathematical symbols. For such examinees, removing the problem stems might preclude an accurate estimate of representation ability. By presenting the stem and the equation, examinees have the choice of using one, the other, or both stimulus components.

A second potential use for the sorting task is as part of a more comprehensive instructional assessment product to help low-scoring examinees localize and remediate difficulties in mathematical problem solving. Such a product might be build around a problem-solving model similar to that of Mayer, Larkin, and Kadane (1984), which hypothesizes four loosely ordered phases: translation, understanding (representation), planning, execution. For example, it is conceivable that some lower-scoring examinees might be able to represent problems accurately but encounter difficulties using those representations to plan and execute solutions. Such examinees would be expected to have problems on constructed-response tests, which have no

multiple-choice options and, thus, no potential cues to execution errors. Other examinees might be adept at execution but deficient in representation and planning. These two patterns imply different instructional actions and, perhaps, different predictions for future scholastic achievement.

Several of the building blocks for such a system exist. Bejar, Embretson, and Mayer (1987) give examples of the types of tasks that might be linked to different problem-solving phases. A basic infrastructure for delivering such problems is the Algebra Assessment System (Sebrechts, Bennett, & Katz, 1993), which allows examinees to enter extended constructed responses to mathematical problems and scores them automatically. Finally, models proposed by Mislevy (1993) and Tatsuoka (1993) might be used to connect item responses to inferences about problem-solving difficulties.

Future research might address several validity issues. First, the current results would be strengthened by evidence showing positive relations with more direct measures of representational skill. For example, examinees might be given an independent set of problems and asked to represent them in the symbol system of their choice: mathematical, diagrammatic, linguistic, or some combination of these. Aside from making a more direct tie to representation, these data might indicate whether the task disadvantages students who tend to represent problems using particular symbol systems. To find out if the proficiencies of mathematically inventive examinees are underestimated, we might interview them to identify the reasons for their categorizations. Finally, particularly critical for any instructional use is whether groups of examinees can be differentiated based on difficulties localized to one or more problem-solving phases and whether targeted instruction improves their performance.

BEST COPY AVAILABLE

References

Bejar, I. I., Embretson, S., & Mayer, R. E. (1987). Cognitive psychology and the SAT: A review of some implications (RR-87-28). Princeton, NJ: Educational Testing Service.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. Cognitive Psychology, 4, 55-81.

Chi, M. T., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. Cognitive Science, 5, 121-152.

Glaser, R. (1991). Expertise and assessment. In M. C. Wittrock & E. L. Baker (Eds.), Testing and cognition. Englewood Cliffs, NJ: Prentice-Hall.

Gliner, G. S. (1989). College students' organization of mathematics word problems in relation to success in problem solving. School Science and Mathematics, 89, 392-404.

Hinsley, D. A., Hayes, J. R., & Simon, H. A. (1977). From words to equations: Meaning and representation in algebra word problems. In M. A. Just & P. A. Carpenter (Eds.), Cognitive processes in comprehension. Hillsdale, NJ: Erlbaum.

Mayer, R. E., Larkin, J. H., & Kadane, J. B. (1984). A cognitive analysis of mathematical problem-solving ability. In R. J. Sternberg (Ed.), Advances in the psychology of human intelligence (pp. 231-273). Hillsdale, NJ: Erlbaum.

Mislevy, R. J. (1993). A framework for studying differences between multiple-choice and free-response test items. In R. E. Bennett & W. C. Ward (Eds.), Construction vs. choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment (pp. 75-106). Hillsdale, NJ: Erlbaum.

Schoenfeld, A. H., & Herrmann, D. J. (1982). Problem perception and knowledge structure in expert and novice mathematical problem solvers. Journal of Experimental Psychology: Learning, Memory, and Cognition, 8, 484-494.

Sebrechts, M. M., Bennett, R. E., & Katz, I. R. (1993). A research platform for interactive performance assessment in graduate education (RR-93-08). Princeton, NJ: Educational Testing Service.

Silver, E. A. (1979). Student perceptions of relatedness among mathematical verbal problems. Journal for Research in Mathematics Education, 10, 195-210.

Silver, E. A. (1981). Recall of mathematical problem information: Solving related problems. Journal for Research in Mathematics Education, 12, 54-64.

BEST COPY AVAILABLE

Tatsuoka, K. K. (1993). Item construction and psychometric models appropriate for constructed responses. In R. E. Bennett & W. C. Ward (Eds.), Construction vs. choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment (pp. 107-133). Hillsdale, NJ: Erlbaum.

Wagner, R. K., Sebrechts, M. M., & Black, J. B. (1985). Tracing the evolution of knowledge structures. Behavior Research Methods, Instruments, & Computers, 17, 275-278.

Young, F. W., & Lewyckyj, R. (1979). ALSCAL-4 user's guide. Carrboro, NC: Data Analysis and Theory Associates.

Appendix A

Sorting Task II Test Directions

GRE Problem Sorting II Test
Directions

Research has shown that the ability to perceive rapidly the underlying mathematical structure of word problems is an important part of problem-solving proficiency. This problem sorting test assesses this ability by asking you to judge the similarity of one word problem's mathematical structure to another's.

If knowledge of a solution method of one problem aids in solving a second problem, the two problems share at least some aspects of mathematical structure. Since any method for solving a word problem involves developing a mathematical model of the problem, two problems would be judged to be structurally similar if there is at least one mathematical model that can be used to solve both problems. Two problems would be judged to be structurally identical if every mathematical model that can be used to solve one of them can be used to solve the other. Note that such surface characteristics as language, story context, and the units used are not necessarily related to mathematical structure.

You will be asked to sort 20 word problems into four groups. Each group is represented by a word problem labeled A, B, C, or D. Each of the four problem statements is followed by one possible algebraic model of it. Each model contains an equation with one variable, x, and several constants, a, b, c, and d (although not every constant appears in each model). Values taken from the problem statement are given for the constants after each equation. You must match each problem with the representative whose given equation can also be used to model the problem being matched, provided values for the constants are taken from the problem statement. When you are finished, each of the four groups should have at least one problem in it, but the final numbers of problems in the groups need not be the same.

For example, consider the following sample word problems representing two different groups.

A. Julia projects that the market price of her house will increase by 3.5% per year for the next 5 years. If the house's market price is currently \$88,000, what does Julia project the market price of her house will be 5 years from now?

   $x = a\,(1 + b)^c$, where $a = 88{,}000$, $b = 0.035$, and $c = 5$.

B. A certain cake recipe requires 5 cups of an ingredient and each full jar of the ingredient contains 11 ounces. If 8 ounces of the ingredient are required to fill one cup, how many full jars of the ingredient must be opened to make the cake?

   $x = \left\lceil \frac{ab}{c} \right\rceil$, where $a = 5$, $b = 8$, and $c = 11$ and $\lceil z \rceil$ means the least integer that is greater than or equal to z.

Also, consider the following word problem from a hypothetical set of 20 problems.

1. On July 1, 1992, Ms. Fox deposited $10,000 in a new account at the annual interest rate of 6% compounded annually. If no additional deposits or withdrawals are made and interest is credited each June 30, how much money will be in the account on July 1, 2007, 15 years after the initial deposit was made?

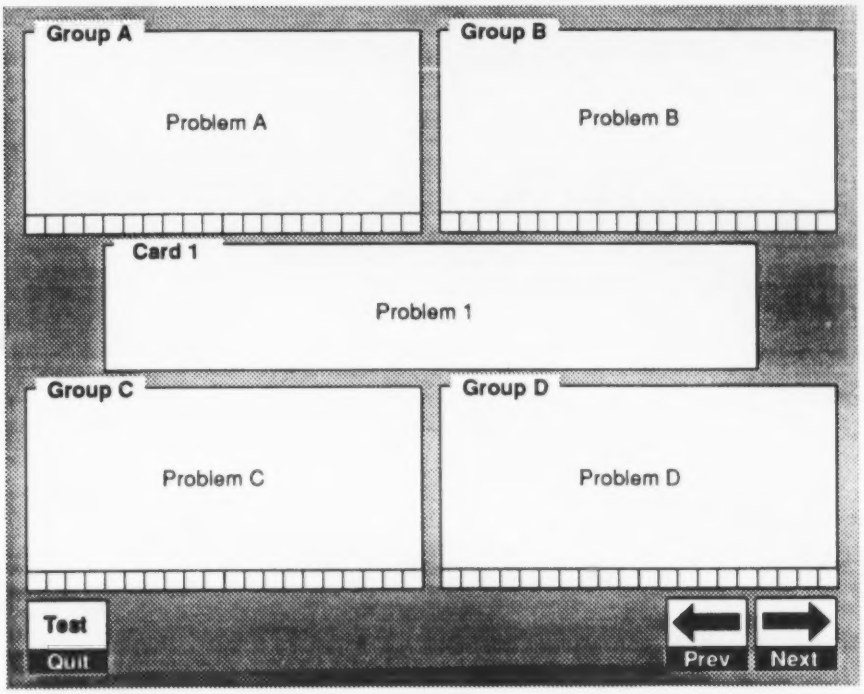
The underlying structure of Problem 1 is similar to that of Problem A since both problems ask about the quantity that results from repeatedly increasing a given quantity by a fixed ratio or percent. Therefore, the equation $x = a\,(1 + b)^c$ can be used to model both problems. (For Problem 1, the values for the constants are $a = 10{,}000$, $b = 0.06$, and $c = 15$.)

Problem B, on the other hand, asks about the number of smaller quantities it would take to obtain a large quantity. The large quantity, in turn, results from repeated additions of another small quantity. Problem 1 does not ask about this. Therefore, the equation $x = \left\lceil \frac{ab}{c} \right\rceil$, where $\lceil z \rceil$ means the least integer that is greater than or equal to z is not a model for this problem. Thus, problem 1 would be matched with problem A as opposed to problem B.

Try to work out each problem with paper and pencil to the degree needed to identify the match and use your notes in comparing the problems. It is best NOT to solve each problem completely because time is limited.

BEST COPY AVAILABLE

GRE Problem Sorting II Test
Directions



Think of this as a card-sorting task with 20 cards. In the center of the screen will be the deck of 20 cards – each card contains one word problem. Sort each card into one of the four groups labeled A, B, C, and D by clicking anywhere within the box representing the group – the number of the card that you just sorted will appear in the row at the lower edge of that box.

You can move forward through the cards by clicking on the NEXT icon or return to a previous card by clicking on the PREVIOUS icon. To change an answer by moving a card from one group to another, first use NEXT or PREVIOUS to bring that card to the screen, then click on the group you want it moved to –it will automatically move to the new group.

Every card must be placed into one of the groups before you can quit the test. When you have finished sorting all the cards and are ready to quit, click on the QUIT icon.

Your score for this test will be based on how many problems you correctly sort into the four groups.

You will have 8 minutes to review the problems and models for groups A, B, C, and D. You will then have 24 minutes to sort the 20 problems.

YOU MUST SORT EACH PROBLEM INTO A GROUP TO RECEIVE PAYMENT
FOR PARTICIPATION IN THIS STUDY.

BEST COPY AVAILABLE

Appendix B

Questionnaire

GRE RESEARCH QUESTIONNAIRE: PROBLEM REPRESENTATION

NAME:_____

SS#:_____

OPINION QUESTIONNAIRE

Please answer each of these questions by circling the letter next to the phrase that best characterizes your opinion. Please remember to answer ALL questions.

1. How easy was the computer-based Problem Representation Test?

    a. Too easy
    b. About right
    c. Too difficult

2. How adequate was the time allowed for answering the computer-based Problem Representation Test?

    a. Too little
    b. About right
    c. Too much

3. Which kind of test question would you rather take: multiple-choice mathematics questions like those found on the SAT and GRE General Test or questions like Problem Representation?

    a. Regular multiple-choice
    b. Problem Representation
    c. No preference

4. Which kind of question do you think is a fairer indicator of your ability to undertake graduate study: multiple-choice mathematics questions like those on the SAT and GRE General Test or questions like Problem Representation?

    a. Regular multiple-choice
    b. Problem Representation
    c. No preference

5. Which kind of test would you rather take: a paper-and-pencil test or a computer-based one?

    a. Paper-and-pencil
    b. Computer-based
    c. No preference

GRE RESEARCH QUESTIONNAIRE: PROBLEM REPRESENTATION

NAME:_____

SS#:_____

6. In the past year, how often have you used a computer?

    a. Never or almost never
    b. About once a week
    c. Daily or almost daily

7. When you have to write a paper for school, how do you usually do it?

    a. Pencil (or pen) and paper
    b. Typewriter
    c. Computer

8. How easy was it to use the computer to answer the Problem Representation questions?

    a. Very easy
    b. Somewhat difficult
    c. Very difficult

9. If you found it "somewhat difficult" or "very difficult" to use the computer, why was that? (Check all that apply.)

    a. The mouse tutorial didn't do a good job explaining how to use the mouse.
    b. The computer screens were confusing.
    c. The sequence of commands was not clear.
    d. The mouse was hard to use.
    e. Other: _____

    _____

GRE RESEARCH QUESTIONNAIRE: PROBLEM REPRESENTATION

NAME:_____

SS#:_____

BACKGROUND QUESTIONS

1. Gender:

    a. Male
    b. Female

2. Undergraduate Major:

    a. Social science
    b. Life science
    c. Physical science
    d. Humanities/arts
    e. Business
    f. Engineering
    g. Education
    h. Other

3. How do you describe yourself?

    a. Asian or Pacific Island American
    b. Black or African-American
    c. Hispanic, Latino, Mexican-American, Puerto Rican, or Central or South American
    d. Native American, American Indian, or Alaskan Native
    e. White (non-Hispanic) or Caucasian
    f. Other

4. Do you plan to apply to graduate school?

    a. Yes
    b. No

5. If **YES**, which of the following major fields will you study?

    a. Social science
    b. Life science
    c. Physical science
    d. Humanities/arts
    e. Business
    f. Engineering
    g. Education
    h. Other

GRE RESEARCH QUESTIONNAIRE: PROBLEM REPRESENTATION

NAME:_____

SS#:_____

6. If you plan to apply to graduate school, which graduate degree will you seek?

   a. Masters degree
   b. Doctoral degree

7. If you have taken the SAT, what was your last math score? _ _ _

   If you have taken the SAT, what was your last verbal score? _ _ _

Additional Comments on the Problem Representation task:

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

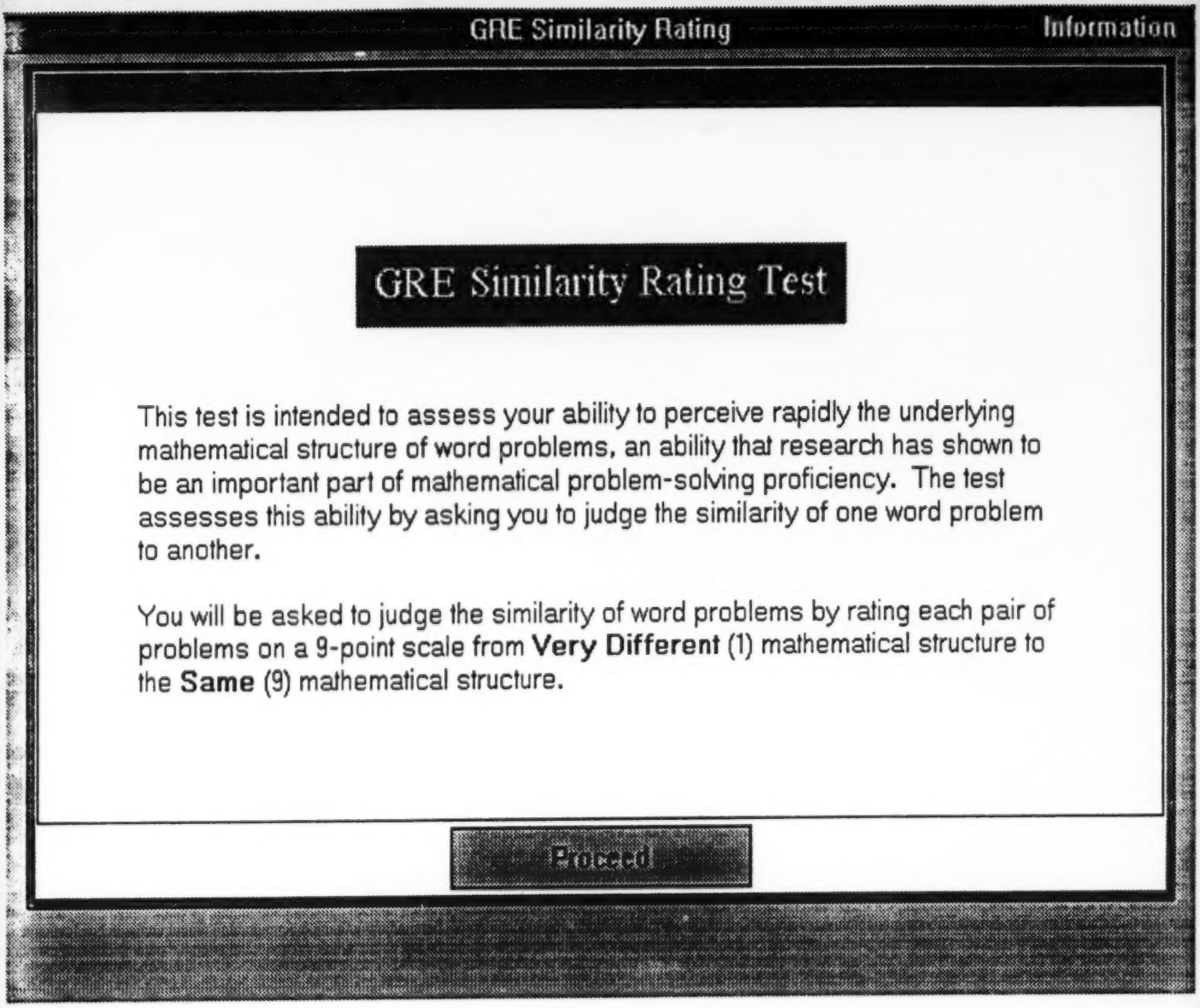
WHEN YOU HAVE COMPLETED THIS FORM, RAISE YOUR HAND TO ALERT THE TEST ADMINISTRATOR THAT YOU ARE THROUGH.

THANK YOU FOR PARTICIPATING IN THIS STUDY!

BEST COPY AVAILABLE

Appendix C

Similarity Rating Task Directions


GRE Similarity Rating
Information
GRE Similarity Rating Test
This test is intended to assess your ability to perceive rapidly the underlying mathematical structure of word problems, an ability that research has shown to be an important part of mathematical problem-solving proficiency. The test assesses this ability by asking you to judge the similarity of one word problem to another.
You will be asked to judge the similarity of word problems by rating each pair of problems on a 9-point scale from **Very Different** (1) mathematical structure to the **Same** (9) mathematical structure.
Proceed

You must select a rating for the comparison of each pair of problems.

On the next screen you will be shown how to:

- locate problems to compare
- rate pairs of problems
- change ratings

Try all of the steps above more than once.

Use the Exit Section icon when finished learning how to rate problems.

Click on Proceed to continue.

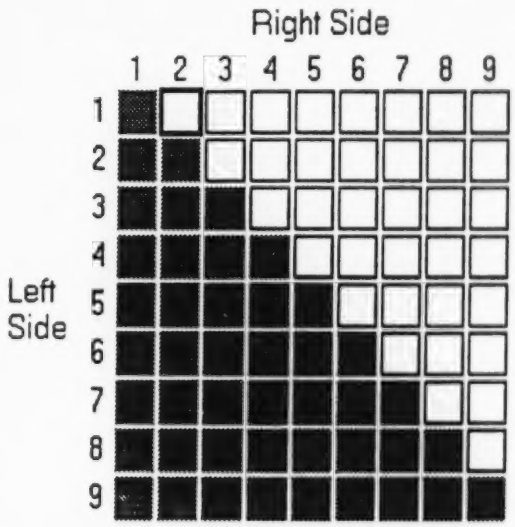
Proceed

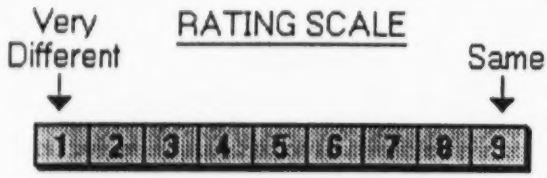
BEST COPY AVAILABLE

GRE Similarity Rating - Tutorial 1 of 1

Item 1 will appear here.

Item 2 will appear here.





To locate a pair of problems, click on the appropriate matrix cell. The cell highlights and the problems appear in the windows at the top left and right.

To rate a pair of problems, click on a number on the scale and the number appears in the selected cell.

To change a rating, click on a matrix cell to highlight it, then click on a new number on the scale.

Section

Exit

GRE Similarity Rating - Tutorial
1 of 1
Exit Section
Once you leave this tutorial, you will not be able to return to it.
If you need more practice,
click on Return to Where I Was.
If you're ready to begin the test,
click on Exit Section.
Return to Where I Was
Exit Section
Section
Exit

GRE Similarity Rating

Directions

Beginning

When finished reading directions click on the icon below

Two problems are considered to have similar mathematical structures to the extent that they share such features as the principle(s) required for solution, the form of the equation(s) that model the problem, the methods that could lead to a solution, the role of variables, and the nature of the steps or operations. If knowledge of the solution to one problem aids in solving a second problem, the two problems share at least some aspects of mathematical structure. Note that such surface characteristics as language, story context, and the units used are NOT necessarily related to mathematical structure.

You will be asked to judge the similarity of word problems by rating each pair of problems on a 9-point scale from **Very Different** (1) mathematical structure to the **Same** (9) mathematical structure. How you rate a given pair of problems should depend on the other items in the set. Therefore, it is a good idea to look briefly over all the problems before you begin your ratings so that you can get a sense of how structurally diverse the problems are. Also, it is a good idea to recheck your ratings periodically to make sure that your judgments are consistent from one pair to the next.

Consider the following set of sample problems.

A. A photocopier enlarges the print of a document to 102 percent of its original size. A document is repeatedly photocopied by photocopying the original and then rephotocopying each successive photocopy.

Dismiss Directions

More Available

the original and then rephotocopying each successive photocopy. What percent of the size of the original print is the print of the final copy if the copier makes five successive photocopies?

B. On July 1, 1992, Ms. Fox deposited $10,000 in a new account at the annual interest rate of 6 percent compounded monthly. If no additional deposits or withdrawals were made and if interest was credited on the last day of each month, what was the amount of money in the account on September 1, 1992 ?

C. A certain cake recipe requires 5 cups of an ingredient and each full jar of the ingredient contains 11 ounces. If 8 ounces of the ingredient are required to fill one cup, how many full jars of the ingredient must be opened to make the cake?

In the context of this 3-problem set, problems A and B have the **Same** (9) underlying structure because they both describe a quantity that is repeatedly increased by a specified ratio or percent, and both ask about the resulting quantity after a number of such increases. The equations used to model these two problems are both of the form $Q = Q_0 r^n$, where Q_0 is the initial quantity, Q is the final quantity, r is the ratio, and n is the number of increases. Q can be computed directly from the given values of Q_0, r, and n. While other methods might be used to model these problems, in all cases the method used to model problem A should work with problem B. Thus, the problems have in common a variety of structural

When finished reading directions click on the icon below

Dismiss Directions

BEST COPY AVAILABLE

GRE Similarity Rating

Directions

More Available

these problems, in all cases the method used to model problem A should work with problem B. Thus, the problems have in common a variety of structural features, including the principle on which a solution is based, the methods used for solution, the role of variables, and the operations needed to achieve the end result.

Problem C is most directly solved by converting from the number of cups needed to the number of ounces needed, and then from ounces to jars. The number of jars is then rounded up to the next whole number in order to answer the question that was asked. The conversions are accomplished by multiplying or dividing by rates or ratios such as 11 ounces per jar. With respect to problems A and B, problem C is structurally quite distant and would be considered **Very Different** (1) from A and B in mathematical structure. Among other things, A and B rely on different principles, methods, and operations than C.

Note that some pairs of problems may not be the **Same** (9) in mathematical structure but may not be **Very Different** (1) either, when compared to other pairs in the set. These intermediary pairs should be rated somewhere between the **Same** (9) and **Very Different** (1). Exactly what rating they should be given depends entirely on how they contrast with other problems in the set. In other words, it is the relativity of ratings that is most important.

In taking the test it is best NOT to solve each problem completely because time is

When finished reading directions click on the box below

Dismiss Directions

GRE Similarity Rating

Directions

End

In taking the test it is best NOT to solve each problem completely because time is limited. Rather, try to work out the problem only to the degree needed to identify its structure. You can then make paper-and-pencil notes and use them in comparing the problem to other problems in the set.

There will be 9 word problems—you will compare each of the 9 problems with the other 8 for the degree of similarity. There is at least one pair of problems in this set that is structurally the **Same** (9) and at least one pair that is structurally **Very Different** (1).

Scores for the task are based on how closely the overall composite of your ratings duplicates the judgments of mathematical experts.

You will have 30 minutes to complete all of the comparisons.

YOU MUST SELECT A RATING FOR EACH COMPARISON TO RECEIVE PAYMENT FOR PARTICIPATION IN THIS PROJECT.

On the next screen you will be able to review all 9 problems before beginning the test.

Click on Dismiss Directions to continue.

Dismiss Directions

BEST COPY AVAILABLE

GRE Similarity Rating Information

Beginning

You will have 5 minutes to scroll through all 9 problems to become familiar with them. Don't waste time trying to solve each one now—you will see them again later. If you finish looking at all 9 problems before the 5 minutes are up, you may begin the test by clicking on Proceed.

1. For a science project, Joyce constructs a scale model of a volcano on a board with the aid of a surveyor's map. The map shows a region with an area of 10 square miles containing the 1-mile-high volcano. If she uses a 5-inch-high can to represent the volcano, how many square inches of the board will be covered by the region shown on the map?

2. Ten children have made model race cars that will be raced two at a time on a two-lane track. If each child's car is to race each of the other cars exactly twice and each race takes about three minutes to hold, including time between races, about how many minutes will it take to hold all of the races?

3. In order to comply with an ordinance concerning residential swimming pools, a man encloses a 50-foot by 70-foot portion of his yard with a brick wall on three sides using a 50-foot side of his house for the remaining side. If the brick wall is likely to cost $15 per foot to build, what will the entire wall likely cost?

4. Twelve fourth graders will march in an Independence Day parade wearing costumes they have made

Proceed

GRE Similarity Rating Information

More Available

4. Twelve fourth graders will march in an Independence Day parade wearing costumes they have made for the occasion. They will parade twice around the edge of a rectangular playing field that is 120 yards long and 50 yards wide. If the children walk 40 yards each minute, about how many minutes will it take them to finish their parade?

5. Jimmy's class must estimate the amount of grass seed that is needed to seed the ground encompassed by their school's 400-meter oval track. They make a similar oval using one meter of string and find that the ground inside that oval requires 1/8 cup of grass seed. About how many cups of seed are needed for the ground inside the track?

6. In order to find grass mixtures that grow well in her lawn, Sue plans to buy four types of grass seed which are sold in trial-size bags. She will plant every possible mixture of two whole bags, one bag of one type and one bag of another, in separately fenced plots in her backyard. If each bag costs $1.35, how much will the seed cost her?

7. In a village where each lot is a square that is 50 feet wide, the zoning board wants to pass a law requiring all fences be set back a certain number of feet from all four property lines. What number of feet should the law stipulate that fences be set back in order to allow owners to fence in 80 percent of the area of their lots?

Proceed

BEST COPY AVAILABLE

GRE Similarity Rating — Information

End

8. An oval track for a miniature train is in the shape of a square with two semicircles attached to opposite sides of the square. If a side of the square is three feet long and a train travels around the track at two feet per second, would it take longer than 10 seconds for the train to travel once around the track?

9. Mr. Smythe's social studies class is comparing the geographies and histories of five South American countries. For each pair of these countries, he has a student give a 10-minute presentation comparing the two countries. Must Mr. Smythe plan for more than two hours of student presentations?

Remember, you'll have 30 minutes to complete all comparisons.
The test and timing will begin when you leave this screen by clicking on Proceed.

Proceed

BEST COPY AVAILABLE